# Estimating the relationship between skill and overconfidence

Maastricht University

Research Centre for Education and the Labour Market | ROA

# Estimating the relationship between skill and overconfidence

Jan Feld
Jan Sauermann
Andries de Grip

## ROA Research Memorandum

ROA-RM-2017/3

Researchcentrum voor Onderwijs en Arbeidsmarkt | ROA
Research Centre for Education and the Labour Market | ROA

# Estimating the relationship between skill and overconfidence

Jan Feld
Jan Sauermann
Andries de Grip

ROA-RM-2017/3*
March 2017

**Research Centre for Education and the Labour Market**
Maastricht University
P.O. Box 616, 6200 MD Maastricht, The Netherlands
T +31 43 3883647 F +31 43 3884914

secretary-roa-sbe@maastrichtuniversity.nl
www.roa.nl

* The ROA Research Memorandum Series was created in order to make research results available for discussion, before those results are submitted for publication in journals.

## Abstract

**Estimating the relationship between skill and overconfidence****

The Dunning–Kruger effect states that low performers vastly overestimate their performance while high performers more accurately assess their performance. Researchers usually interpret this empirical pattern as evidence that the low skilled are vastly overconfident while the high skilled are more accurate in assessing their skill. However, measurement error alone can lead to a negative relationship between performance and overestimation, even if skill and overconfidence are unrelated. To clarify the role of measurement error, we restate the Dunning–Kruger effect in terms of skill and overconfidence. We show that we can correct for bias caused by measurement error with an instrumental variable approach that uses a second performance as instrument. We then estimate the Dunning–Kruger effect in the context of the exam grade predictions of economics students, using their grade point average as an instrument for their exam grade. Our results show that the unskilled are more overconfident than the skilled. However, as we predict in our methodological discussion, this relationship is significantly weaker than ordinary least squares estimates suggest.

JEL classification: D03, I23
Keywords: Dunning-Kruger effect, overconfidence, judgment error, measurement error, instrumental variable

Jan Feld
School of Economics and Finance
Victoria University of Wellington
P.O. Box 600
Wellington 6140
New Zealand
jan.feld@vuw.ac.nz
and IZA

Jan Sauermann
Swedish Institute for Social Research
Stockholm University
Universitetsvägen 10 F (floors 8 and 9)
SE-106 91 Stockholm
Sweden
jan.sauermann@sofi.su.se
and IZA

Andries de Grip
Maastricht University
ROA
P.O. Box 616
NL-6200 MD Maastricht
The Netherlands
a.degrip@maastrichtuniversity.nl
and IZA

** We thank Christian Kerckhoffs and Alexander Vostroknutov for access to their courses. We further thank Adam Booij, Luke Chu, Thomas Dohmen, Jonas Lang, Daniel Pollmann, Ingrid Rohde, Nicolas Salamanca, Stefan Trautmann and Anna Zseleva for valuable comments on earlier drafts of this paper. We gratefully acknowledge financial support from the Network Social Innovation (NSI) at Maastricht University.

## 1. Introduction

Dunning and Kruger (1999) argue that the low skilled are typically vastly overconfident while the high skilled assess their skill more accurately. As evidence for this argument, they present an empirical pattern that is now known as the Dunning–Kruger effect: For many different tasks, low performers typically vastly overestimate their performance while high performers more accurately assess their performance (Dunning, 2011).

This evidence, however, is not sufficient, because measurement error alone can cause low performers to overestimate their performance more than high performers. To understand why, we first need to distinguish between skill and overconfidence and their measures performance and overestimation. Performance is the score on a test and overestimation is the difference between the expected and the actual test score. Performance measures skill with some error. We define skill as the ability to perform well on a given test and we can think of measurement error as luck on this test. Overestimation measures overconfidence, the difference between self-assessed and actual skill.[1]

The source of the bias is that researchers typically use the same performance to measure skill and to calculate overestimation. The same measurement error component is therefore part of performance and overestimation. To see how this can contribute to the Dunning–Kruger effect, consider a person with bad luck on a test: Bad luck decreases performance and increases overestimation and thus makes the person *appear* less skilled and more overconfident.

[1] In their survey of the overconfidence literature, Moore and Healy (2008) define overestimation as overestimation of one's actual performance, overplacement as overestimation of one's performance relative to others, and overprecision as excessive precision in one's beliefs. While these definitions are helpful in distinguishing between the different domains of overconfidence, they are all defined in terms of actual outcomes, which may be affected by measurement error. We thus follow Moore and Healy's definition of overestimation and additionally define overconfidence as the overestimation of one's skill.

In the methodological part of this paper, we discuss the role of measurement error in the estimation of the Dunning–Kruger effect. We first restate the effect in terms of skill and overconfidence instead of their measures. We then show how measurement error causes an overestimation of the Dunning–Kruger effect and the assumptions under which we can correct for this bias with an instrumental variable (IV) approach.

In the empirical part of this paper, we estimate the Dunning–Kruger effect with a sample of economics students who we asked four weeks before the exam to predict their exam performance. In line with the previous literature, we find that students who performed poorly on the exam also vastly overestimated their exam performance. We then estimate the Dunning–Kruger effect with an IV approach, using students' first-year grade point average (GPA) as an instrument for their exam performance. Our results confirm that the effect exists: The low skilled are vastly overconfident and the high skilled are more accurate in assessing their skill. As predicted by our methodological discussion, this effect is, however, significantly smaller than ordinary least squares (OLS) estimates suggest.

Krueger and Mueller (2002) are the first to have pointed out that measurement error can cause bias in the estimation of the Dunning–Kruger effect.[2] They correct for this bias by using two test performances: one to measure skill and one to calculate overestimation. They then regress overestimation calculated with the first performance on the second performance. The advantage of this approach is that it breaks the mechanical relationship between performance and overestimation, because the measurement error parts are now different for both variables. The disadvantage of this approach, however, is that the measurement error of the second performance (the independent variable) may bias the estimates toward zero. Low

[2] Krueger and Mueller (2002) argue that the Dunning–Kruger effect may be a statistical artifact caused by regression effects and the better-than-average effect. Their argument is that regression effects would lead to equal overestimation for low performers and underestimation for high performers. However, the fact that people are generally overconfident leads to an increase in the overestimation of the low performers and a decrease in the underestimation of the high performers. These two forces together therefore lead to the high overestimation of the low performers and the accurate performance assessment of the high performers.

test–retest correlations of the test performances used by Krueger and Mueller (2002) suggest this measurement error is substantial (the test–retest correlation is 0.17 for their difficult test and 0.56 for their easy test). This could be why Krueger and Mueller do not find evidence of the Dunning–Kruger effect.

In response to Krueger and Mueller (2002), Ehrlinger et al. (2008) estimate the Dunning–Kruger effect using reliability-adjusted OLS. They regress overestimation on performance and then divide the estimated performance coefficient by a measure of the test's reliability. They thus present evidence of the Dunning–Kruger effect. Their approach is, however, problematic, because they still use the same performance as a measure of skill and to calculate overestimation. The performance coefficient of this regression is therefore likely biased and adjusting for test reliability only increases this bias.[3]

## 2. Dunning–Kruger Effect

The setup of Dunning–Kruger effect studies is straightforward. Participants take a test in a given domain (e.g., English grammar, understanding humor, gun safety knowledge) and guess their performance on this test either before or after the test. The main finding is that bottom quartile performers vastly overestimate their performance while top quartile performers more accurately assess their performance.[4] This finding has been widely replicated with different populations and for a number of different tasks (Ehrlinger et al., 2008; Ryvkin, Krajč, & Ortmann, 2012; Schlösser, Dunning, Johnson, & Kruger, 2013).

Dunning and Kruger (1999) interpret this finding as evidence of a negative relationship between skill and overconfidence as opposed to merely an empirical pattern, which can be seen from the title of their paper: "Unskilled and Unaware of It: How

[3] See Feld (2014) for a more extensive discussion on the biases of other estimation methods.
[4] When using relative performance measures, high-performing individuals typically slightly underestimate their performance. Kruger and Dunning (1999) explain this with the false consensus effect (Ross, Greene, & House, 1977), which states that people tend to overestimate the degree to which people are similar to them. The high-skilled overestimate the performance of others and therefore slightly underestimate their relative performance.

Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments." They further argue that differences in metacognitive skills between the low and high skilled drive this relationship. The idea is that the skills necessary to perform well are also those necessary to evaluate one's performance accurately. The low skilled therefore perform badly and lack the metacognitive skills to realize it.[5] As evidence for this explanation, Dunning and Kruger show that a randomly assigned training can increase competence and decrease the overestimation of low performers.

## 3. Estimating the Dunning–Kruger Effect

### 3.1. Key Variables

We define skill broadly as the ability to perform well on a given test. Skill, however, is imperfectly measured by performance, which is partly determined by luck. We use the terms *luck* and *measurement error* interchangeably to refer to all factors besides skill that influence test performance. We therefore define performance $p$ as the sum of skill $s^*$ and a classical measurement error component $\varepsilon$ (asterisks indicate unobserved variables and we omit individual subscripts to simplify notation):

$$p \equiv s^* + \varepsilon. \qquad (1)$$

Classical measurement error means that $\varepsilon$ is a random error term with a mean of zero and independent of all variables included in the regression and the error term.

[5] Krajč and Ortmann (2008) propose an alternative explanation for the Dunning–Kruger effect. They observe that many of the studies showing the Dunning–Kruger effect use students from very selective institutions and argue that the students' skills in these samples follow a J-distribution equivalent to the upper tail of a normal distribution. The authors then show that the Dunning–Kruger pattern can arise even if people make random judgment errors, due to the J-distribution of skills and floor and ceiling effects caused by the test scale. In response, Schlösser et al. (2009) argue that, because student admission is based on many criteria, even in very selective institutions, skill is likely close to normally distributed. They then show that, even in the rare cases where skill follows a J-distribution, the Krajč–Ortmann explanation would only account for a small fraction of the observed Dunning–Kruger effect.

We define overconfidence as the difference between self-assessed skill and actual skill, that is, $oc^* \equiv s^*_{self-assessed} - s^*$. Overconfidence, however, is imperfectly measured by overestimation, that is, the difference between expected and actual performance:

$$oe \equiv p_{exp} - p. \tag{2}$$

We further assume that people state their self-assessed skill when asked about their expected performance $p_{exp}$. Expected performance is therefore the sum of a person's actual skill and overconfidence:[6]

$$p_{exp} \equiv s^* + oc^* = s^*_{self-assessed} \tag{3}$$

When we decompose overestimation into its respective elements, shown in Equations (1) and (3), we can see that it is equal to overconfidence minus measurement error:

$$\begin{aligned} oe &= (s^* + oc^*) - (s^* + \varepsilon), \\ oe &= oc^* - \varepsilon. \end{aligned} \tag{4}$$

We can see from Equations (1) and (4) that the same measurement error component is part of performance and overestimation. This is the source of the OLS bias. Intuitively, we can see that bad luck (i.e., negative $\varepsilon$) decreases performance and increases overestimation and thus make the test taker *appear* less skilled and, at the same time, more overconfident.

### 3.2. Restating the Dunning–Kruger Effect in Terms of Skill and Overconfidence

We model overconfidence $oc^*$ as a linear function of skill $s^*$ and an error term *u* that captures individual differences in overconfidence unrelated to skill:

$$oc^* = \alpha + \beta_1 s^* + u. \tag{5}$$

Equation (5) provides a simple framework to restate the Dunning–Kruger effect in terms of skill and overconfidence. The (restated) Dunning–Kruger effect is that overconfidence among

[6] Besides expected skill, a number of other factors might influence a person's expected performance. When expected performance is elicited before the test, as in this paper, these other factors are arguably unrelated to skill and measurement error and therefore do not affect Dunning–Kruger effect estimates.

the low skilled is large and positive ($\alpha + \beta_1 s^*$ is large and positive for low values of $s^*$) while overconfidence among the high skilled is small in absolute value ($\alpha + \beta_1 s^*$ is small in absolute value for high values of $s^*$). This effect implies that skill is negatively related to overconfidence, which means that $\beta_1$ is negative. To focus our discussion on the role of measurement error, we assume throughout that the error term $u$ has a mean of zero and is independent of all included variables.

### 3.3. OLS Bias

Researchers typically estimate the Dunning–Kruger effect by either regressing overestimation on performance or by showing the average overestimation by performance quartile (Kruger & Dunning, 1999; Burson, Larrick, & Klayman, 2006; Ehrlinger, et al., 2008; Ryvkin, et al., 2012). Both approaches result in biased estimates for similar reasons and we focus on showing the bias in a regression framework.

To show the OLS bias, we express Equation (5) in terms of observable variables. It follows from Equations (1) and (4) that $oc^* = oe + \varepsilon$ and $s^* = p - \varepsilon$. When we substitute these into Equation (5) and rearrange, we obtain the following expression:

$$\begin{aligned} oe &= \alpha + \beta_1 p + u - \varepsilon(1 + \beta_1), \\ oe &= \alpha + \beta_1 p + \omega, \end{aligned} \tag{6}$$

which shows that simply regressing overestimation on performance leads to biased estimates of $\beta_1$ because the luck component of performance is also part of the composite error term $\omega = u - \varepsilon(1 + \beta_1)$.

We can then show the direction of the bias with the formula for the bias of the simple OLS slope estimator, $\frac{Cov(p,\omega)}{Var(p)}$. When we decompose and rearrange this bias, we can rewrite it as

$$-(1+\beta_1)\frac{Var(\varepsilon)}{Var(p)}. \quad (7)$$

Equation (7) shows that the sign of the bias depends on $\beta_1$. We expect $\beta_1 >$ -1 because $\beta_1 \leq$ -1 would mean that self-assessed skill stays constant or even *declines* with actual skill. In addition, $\beta_1 \leq$ -1 means that a one-point increase in skill causes overconfidence to decrease by at least one point. Because self-assessed skill is simply the sum of skill and overconfidence (see Eq. (3)), this would mean that, as skill increases, a person's self-assessed skill would stay constant (if $\beta_1$ = -1) or even decrease (if $\beta_1 <$ -1), an unrealistic scenario. If $\beta_1 >$ -1, it is straightforward to see that the OLS bias is negative. We therefore expect OLS estimates to be more negative than $\beta_1$ (i.e., larger in absolute terms) and therefore overestimate the Dunning–Kruger effect.

The size of the bias depends on the true relationship between skill and overconfidence and the degree of measurement error. To obtain an idea about its potential magnitude, consider, for example, that there is no relationship between skill and overconfidence (i.e., $\beta_1 = 0$) and the ratio of measurement error variance to performance variance is one-half (i.e., $\frac{Var(\varepsilon)}{Var(p)} = 0.5$).[7] We can see from Equation (7) that the OLS bias is -0.5. In this example, OLS estimates, on average, show that a one-point increase in performance is associated with a decrease of 0.5 in overestimation, even though skill and overconfidence are unrelated.

### 3.4. Correcting for Bias Caused by Measurement Error with IVs

What do we mean by correcting for bias caused by measurement error? Imagine a situation in which we could perfectly measure skill and overconfidence. In this situation, we could simply regress overconfidence on skill without worrying about measurement error. The following equation shows the skill coefficient of such a regression:

[7] A standard way to quantify the degree of measurement error is test reliability, which is defined as $Var(s^*)/Var(p)$. If measurement error is random, $Var(s^*)/Var(p) + Var(\varepsilon)/Var(p) = 1$ and, so, $Var(\varepsilon)/Var(p)$ is simply one minus test reliability.

$$\frac{Cov(oc^*, s^*)}{Var(s^*)}. \quad (8)$$

This skill regression coefficient is our benchmark. We say that IV corrects for measurement error if the IV estimator of $\beta_1$ is equal to Equation (8).

To use the IV method, we need a second performance, $p_2$, as an IV:

$$p_2 \equiv s^* + \varepsilon_2, \quad (9)$$

where $\varepsilon_2$ is the measurement error term. The IV estimator of $\beta_1$ is then equal to $\frac{Cov(oe,p_2)}{Cov(p,p_2)}$.

We can decompose the IV estimator to show the assumptions under which it corrects for measurement error:

$$\frac{Cov(oc^*, s^*) + Cov(oc^*, \varepsilon_2) - Cov(\varepsilon, s^*) - Cov(\varepsilon, \varepsilon_2))}{Var(s^*) + Cov(s^*, \varepsilon_2) + Cov(\varepsilon, s^*) + Cov(\varepsilon, \varepsilon_2))}. \quad (10)$$

We can see from Equation (10) that, if luck in both skill measures is random, the IV estimator is equal to the skill regression coefficient shown in Equation (8), because all the covariances with the measurement error terms are equal to zero. More specifically, we assume that the original performance measures skill with classical measurement error, which implies that the measurement error component of the original performance is unrelated to skill ($Cov(\varepsilon, s^*) = 0$). If this is the case, the IV estimator corrects for measurement error if the measurement error of the second performance (the instrument) is uncorrelated with overconfidence ($Cov(oc^*, \varepsilon_2) = 0$), skill ($Cov(s^*, \varepsilon_2) = 0$), and the measurement error of the original performance ($Cov(\varepsilon, \varepsilon_2) = 0$).

### 3.5. Eliciting Performance Expectations after the Test

So far, we have discussed the estimation bias when expected performance is elicited before instead of after the test. This approach simplifies the discussion of the estimation bias and matches our empirical application. In many studies, however, researchers elicit performance expectations after the test (e.g., Kruger and Dunning, 1999; Ehrlinger et al., 2008). We

extensively discuss how this difference in study design affects estimates of the Dunning–Kruger effect in Section A1 of the Appendix. The busy reader can skip this discussion and take away three key points. First, the empirical relationship between skill and overconfidence may be different after the test. Having taken a test provides feedback about one's skill and Ryvkin, Krajč, and Ortmann (2012) have shown that feedback improves calibration, particularly among the low-skilled. This finding suggests that the Dunning–Kruger effect is less pronounced after the test. Second, the test taker may know part of his or her luck after the test. Accounting for this luck when stating expected performance can decrease the estimation bias. However, OLS estimates are still biased and potentially overestimate the Dunning–Kruger effect if at least part of the test luck is still unknown after the test. Third, we can still correct for bias caused by measurement error with IV estimation. All we need is a second performance as an instrument, as long as the measurement error of this performance is uncorrelated with skill, overconfidence, and the known and unknown luck on the test.

## 4. Data

We estimate the Dunning–Kruger effect with a sample of 89 economics students of a second-year bachelor course at the School of Business and Economics of Maastricht University, in the Netherlands.[8] The course was given in March and April 2013. A total of 94 percent of the students in our sample were in the same bachelor of economics program and the course is compulsory for specialization in this program. The remaining 6 percent took the course as an elective. In total, 75 (84 percent) students filled out the questionnaire. The remaining 14 students were not present on the day the questionnaire was distributed in the classroom, either because they missed the particular session or had already dropped out of the course. Because Maastricht is close to the German border, the School of Business and Economics has a large

[8] See Feld, Salamanca, and Hamermesh (2016) for more information on the school's institutional background.

share of German students. In our estimation sample, 48 percent of students are German and 30 percent are Dutch; 28 percent are female.[9]

We elicited students' predictions of their exam grades with a questionnaire four weeks before the exam. Grades are given on a scale from zero (lowest) to 10 (highest). The minimal exam grade necessary to pass the course is 5.5. To encourage students to state their honest expectations, we incentivized the exam grade predictions by holding a lottery draw in which students could win one of two gift vouchers worth €20 if their prediction was within 0.25 points of their actual exam grade (see the questionnaire in the Appendix). Furthermore, the students were assured that all information would be kept confidential. Information on actual grades was provided by the course coordinators; information on student characteristics and previous grades was taken from the administrative records. The final estimation sample comprises 67 students due to missing data on final grades and GPAs.

Table 1 shows the summary statistics for the estimation sample of students' predictions, the actual grades, the resulting over- and underestimation, and the students' GPAs at the end of the first year, which consists of eight different grades for the typical student.[10] On average, students significantly overestimated their exam grades by 0.63 grade points ($p$ = 0.004). Figure 1 shows the distribution of exam grades.

[9] We also collected similar data from another course, which we do not use in this paper because bunching of grades at the highest possible exam grade for this course makes the classical measurement error assumption unrealistic. Including this course hardly changes our estimates. We furthermore elicited students' expectations about the percentile of their exam grades and their participation grades. We do not use students' percentile expectations because the grade percentile is a relative performance measure and the classical measurement error assumption is therefore unrealistic. We do not use the participation grade predictions to test the Dunning–Kruger effect because we do not have a suitable instrument for the participation grade.

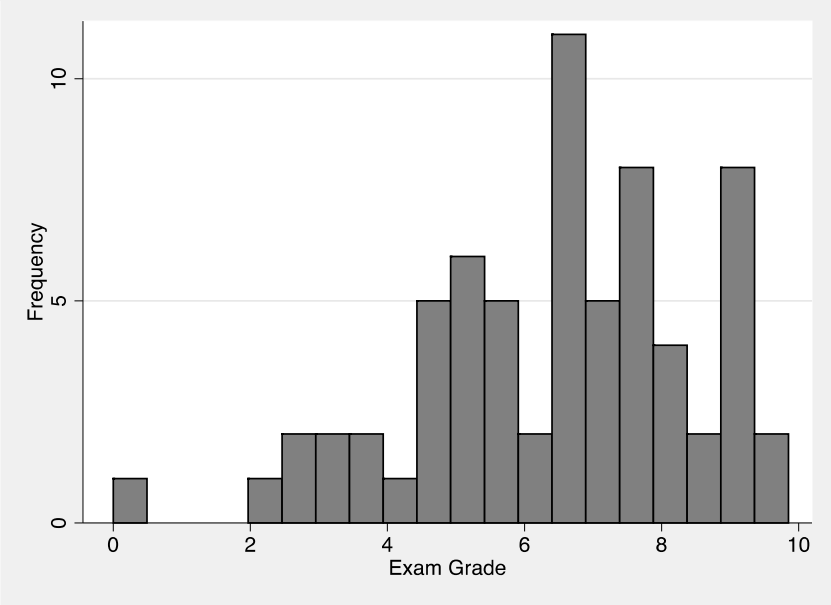
[10] Note that only one student has the lowest possible exam grade and no student has the highest possible exam grade, which shows that no floor or ceiling effects are caused by the grade scale. The GPA is a weighted average–by the European Credit Transfer and Accumulation System (ECTS) course credit points—of all the graded components available at the end of the academic year 2011/2012. The same data are used by Feld and Zölitz (2017). For most of the students, the GPA measure consisted of eight regular courses (6.5 ECTS) and two skills courses (three ECTS) that are compulsory in the first year of the bachelor of economics program.

**Table 1: Predictions, Grades, and Overestimation**

| | Mean | S.D. | Min | 0.25 | 0.50 | 0.75 | Max |
|---|---|---|---|---|---|---|---|
| Expected exam grade | 7.07 | 0.86 | 5.50 | 6.50 | 7.00 | 7.50 | 9.25 |
| Realized exam grade | 6.45 | 2.01 | 0.00 | 5.10 | 6.65 | 7.80 | 9.85 |
| Exam overestimation | 0.63 | 1.70 | -2.55 | -0.60 | 0.25 | 1.65 | 5.50 |
| GPA | 7.11 | 1.41 | 4.04 | 5.97 | 7.42 | 8.33 | 9.38 |

**Note:** The data in this table are based on the estimation sample. Exam overestimation is equal to the expected exam grade minus the realized exam grade.

**Figure 1: Distribution of Exam Grades**

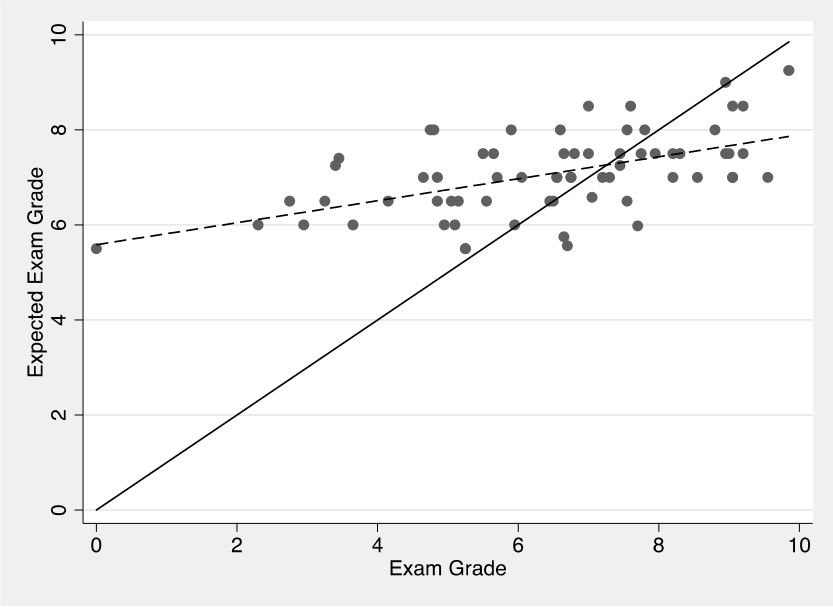



**Note:** This figure is based on the estimation sample.

## 5. Results

Figure 2 plots the exam predictions against the actual exam grades. If all students had perfect foresight about their exam grades, the relationship between the predicted and actual grades would be shown by the 45-degree (solid) line. The figure shows the typical pattern of many Dunning–Kruger effect studies: Those with low grades vastly overestimate their grades while those with high grades, on average, slightly underestimate them. However, as discussed in Section 3, the relationship between performance (actual grades) and overestimation shown in Figure 1 is at least partly caused by measurement error.

**Figure 2: Actual versus Expected Exam Grades**




**Note:** This figure is based on the estimation sample. The solid line is the 45-degree line and the dashed line is the OLS regression line.

To correct for bias caused by measurement error, we estimate the Dunning–Kruger effect using a two-stage least squares IV approach. The following equations show the first and second stages:

$$grade = \zeta_0 + \zeta_1 \, GPA + \theta \tag{11}$$

$$overestimation = \delta_0 + \delta_1 \, \widehat{grade} + \eta, \tag{12}$$

where the variable names are self-explanatory.

The IV estimation corrects for bias caused by measurement error if the instrument GPA fulfills two assumptions. First, it needs to be correlated with the exam grade. This is plausible because similar skills often determine grades in different courses. Second, the GPA measurement error needs to be uncorrelated with skill, overconfidence, and the exam grade's measurement error (see Section 3.4). We think of measurement error as mostly transitory and unpredictable factors that influence academic performance, such as guessing the correct answer on multiple choice questions and disturbing seat neighbors during an exam.[11] These

[11] Recall that we define skill as the ability to perform well on a test. Factors that are predictable and consistently influence test performance are therefore included in our broad definition of skill.

factors are arguably unrelated to skill and overconfidence. We might be worried that the measurement error components of the exam and GPA are correlated, because some factors, such as being sick, affect students' performance in multiple time periods. However, we do not think this is a concern because the last grade of the GPA was graded eight months before the exam. The GPA measurement error is therefore likely uncorrelated with the measurement error of the exam.

Table 2 shows estimates of the Dunning–Kruger effect. We report OLS estimates in Column (1) as a benchmark. The OLS estimate shows that a one-point increase in the exam grade is associated with a decrease of 0.77 in overestimation. This estimate, however, is likely too large (in absolute terms) because of measurement error. Column (2) shows the first stage of the IV estimation. As expected, the GPA is highly predictive of a student's exam grades. The *F*-statistic of the excluded instrument is large, which means that we do not worry about weak instrument bias. Column (3) shows the estimated coefficients of the second stage. The estimated effect of skill is negative and highly significant. An increase in skill of one grade point is related to a decrease in overconfidence by 0.60 grade points, a large effect that is substantially smaller (i.e., less negative) than OLS estimates would suggest. The Wu–Hausmann test shows that the difference between OLS and IV estimates is statistically significant (*p*-value: 0.004). This result confirms that measurement error causes a substantial overestimation of the Dunning–Kruger effect. Columns (4) to (6) show that including additional controls for student characteristics hardly changes the OLS or IV estimates and also with additional controls the Wu-Hausmann test confirms that both estimates are significantly different from each other (p-value: 0.013).

**Table 2: Estimates of the Dunning–Kruger Effect**

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | OLS | First Stage | Second Stage | OLS | First Stage | Second Stage |
| Dep. Variable | Overest. | Exam Grade | Overest. | Overest. | Exam Grade | Overest. |
| Exam grade | -0.769*** | | -0.600*** | -0.781*** | | -0.619*** |
| | (0.039) | | (0.073) | (0.034) | | (0.077) |
| GPA | | 0.855*** | | | 0.799*** | |
| | | (0.152) | | | (0.170) | |
| Constant | 5.583*** | 0.370 | 4.495*** | 6.249*** | 0.553 | 5.211*** |
| | (0.258) | (1.155) | (0.479) | (0.366) | (1.492) | (0.581) |
| Controls | No | No | No | Yes | Yes | Yes |
| F excl. instrument | | 31.6 | | | 22.0 | |
| Observations | 67 | 67 | 67 | 67 | 67 | 67 |
| R-squared | 0.821 | 0.362 | 0.782 | 0.847 | 0.394 | 0.815 |

**Note:** Robust standard errors are in parentheses. Additional controls include dummy variables for female, German, Dutch, and field of study (economics = 1). *** p < 0.01, ** p < 0.05, * p < 0.1.

We might be worried that the IV estimates are biased if the GPA measurement error has a direct effect on overconfidence. Having good luck in the first year may cause students to be more overconfident. This mechanism would lead to a positive correlation between the GPA measurement error and overconfidence and thus a positive bias of the estimate of the relationship between skill and overconfidence.[12] We think such a bias is, if it exists, small for two reasons. First, the measurement error component of the GPA is arguably small because the GPA is the average of the grades of eight first-year courses. Second, the effect of GPA measurement error on overconfidence is, if anything, small, because the GPA consists of courses that the students took at least eight months before the exam. Further, when we compare this potential bias with the OLS bias, we can see that the biases go in different directions. OLS estimates are negatively biased and therefore the Dunning–Kruger effect is overestimated. This bias means that, even though we observe a negative and statistically significant OLS coefficient, we cannot rule out that there is no relationship between skill and

[12] This bias can also be seen in Equation (8) where a positive effect of luck component on GPA on overconfidence is reflected in $Cov(oc^*, \varepsilon_2) > 0$.

overconfidence. In contrast, the potential IV bias would lead to an underestimation of the Dunning–Kruger effect. We therefore interpret our estimates as a lower bound.

Overall, our results provide evidence of the Dunning–Kruger effect: The negative coefficient of the (predicted) exam grade shows that overconfidence declines with skill. We can further use the predicted exam grades, our unbiased measure of skill, and the estimates of $\alpha$ and $\beta_1$ in Column (3) of Table 2 to demonstrate that the Dunning–Kruger effect holds in our sample: The predicted overconfidence of the student of the 10th percentile of the skill distribution (predicted exam grade = 4.49) is equal to 1.41 (4.68 - 0.60*4.49) while the predicted overconfidence of a student of the 90th percentile of the skill distribution (predicted exam grade = 7.83) is equal to -0.20 (4.68 - 0.60*7.83). In line with the Dunning–Kruger effect, low-skilled students are very overconfident while high-skilled students are more accurate in assessing their skill.

## 6. Conclusion

We have shown how measurement error can lead to a negative relationship between performance and overestimation, even if skill and overconfidence are unrelated. We have estimated the Dunning–Kruger effect using an IV approach. Our findings support the existence of the Dunning–Kruger effect: Low-skilled students are very overconfident while high-skilled students are more accurate in assessing their skill. As expected from our methodological discussion, this relationship is significantly weaker than OLS estimates would suggest. This result confirms that taking measurement error into account is crucial when estimating the Dunning–Kruger effect.

**References**

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or Unskilled, But Still Unaware of it: How Perceptions of Difficulty Drive Miscalibration in Relative Comparisons. *Journal of Personality and Social Psychology, 90*(1), 60-77.

Dunning, D. (2011). The Dunning-Kruger Effect: On Being Ignorant of One's Own Ignorance. *Advances in Experimental Social Psychology, 44*, 247-296.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the Unskilled Are Unaware: Further Explorations of (Absent) Self-insight Among the Incompetent. *Organizational Behavior and Human Decision Processes, 105*(1), 98-121.

Feld, J. (2014). Unskilled and Unaware? On Estimating the Relationship Between Skill and Overconfidence, *Making the Invisible Visible - Essays on Overconfidence, Discrimination and Peer Effects*. Maastricht.

Feld, J., Salamanca, N., & Hamermesh, D. S. (2016). Endophilia or Exophobia: Beyond Discrimination. *The Economic Journal*. 126, 1503-1527.

Feld, J., & Zölitz, U. (2017). Understanding Peer Effects: On the Nature, Estimation and Channels of Peer Effects. *Journal of Labor Economics*, 35(2).

Krajč, M., & Ortmann, A. (2008). Are the Unskilled Really That Unaware? An Alternative Explanation. *Journal of Economic Psychology, 29*(5), 724-738.

Krueger, J., & Mueller, R. A. (2002). Unskilled, Unaware, or Both? The Better-than-average Heuristic and Statistical Regression Predict Errors in Estimates of Own Performance. *Journal of Personality and Social Psychology, 82*(2), 180-188.

Kruger, J., & Dunning, D. (1999). Unskilled and Unaware of it: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121-1134.

Kruger, J., & Dunning, D. (2002). Unskilled and Unaware--but Why? A Reply to Krueger and Mueller (2002). *Journal of Personality and Social Psychology*, *82*(2), 189-192.

Moore, D. A., & Healy, P. J. (2008). The Trouble with Overconfidence. *Psychological Review*, *115*(2), 502-517.

Ross, L., Greene, D., & House, P. (1977). The "False Consensus Effect": An Egocentric Bias in Social Perception and Attribution Processes. *Journal of Experimental Social Psychology*, *13*(3), 279-301.

Ryvkin, D., Krajč, M., & Ortmann, A. (2012). Are the Unskilled Doomed to Remain Unaware? *Journal of Economic Psychology, 33*(5), 1012-1031.

Schlösser, T., Dunning, D., Johnson, K. L., & Kruger, J. (2013). How Unaware are the Unskilled? Empirical Tests of the "Signal Extraction" Counterexplanation for the Dunning-Kruger Effect in Self-Evaluation of Performance. *Journal of Economic Psychology*, 39, 85-100.

## APPENDIX

### A1: Eliciting Performance Expectations after the Test

The relationship between skill and overconfidence may be different before and after the test. To allow for this possibility, we model overconfidence *after* the test, $oc_a^*$, as a linear function of skill:

$$oc_a^* = \gamma_0 + \gamma_1 s^* + \tau. \quad \text{(A1)}$$

The main difference between this model and Equation (4) is that we allow the effect of skill on overconfidence to be different after the test. Analogously to the discussion in Section 3.2, we assume that the error term $\tau$ has a mean of zero and is independent of all included variables. Taking a test provides feedback on one's skill and Ryvkin, Krajč, and Ortmann (2012) show that feedback improves calibration, particularly among low performers. We therefore expect the relationship between skill and overconfidence to be less pronounced after the test, that is, $\gamma_1$ is smaller in absolute terms (i.e., less negative) than $\beta_1$.

The main difference in estimating the Dunning–Kruger effect is that luck on the test may be known—at least partly—after the test. To allow for luck that is known and luck that is unknown after the test, we express performance as

$$p = s^* + \varepsilon_k + \varepsilon_u, \quad \text{(A2)}$$

where $\varepsilon_k$ and $\varepsilon_u$ are the known and unknown parts of luck, which we assume to be random mean-zero error terms. In particular, we assume that both error terms are uncorrelated with skill overconfidence and each other. To obtain an idea about both types of luck, consider a student taking a test. The student likely knows after the test whether he or she prepared for the right kind of questions, but not whether he or she guessed correctly on multiple choice questions. We further assume that students account for their known luck but not for their unknown luck when stating their expected performance after the test, $p_{exp-a}$:

$$p_{exp-a} \equiv s^* + oc_a^* + \varepsilon_k \quad \text{(A3)}$$

We define overestimation after the test, $oe_a$, as the difference between expected performance after the test and performance,

$$oe_a \equiv p_{exp-a} - p, \tag{A4}$$

which can be decomposed into its respective elements shown in Equations (A2) and (A3) to show that overestimation after the test is equal to overconfidence after the test minus unknown luck:

$$\begin{gathered} oe_a = (s^* + oc_a^* + \varepsilon_k) - (s^* + \varepsilon_k + \varepsilon_u) \\ oe_a = oc_a^* - \varepsilon_u. \end{gathered} \tag{A4}$$

We then can express $s^* = p - \varepsilon_k - \varepsilon_u$ and $oc_a^* = oe_a + \varepsilon_u$. If we substitute and rearrange these into Equation (A1), we obtain the following expression:

$$\begin{gathered} oe_a = \gamma_0 + \gamma_1 p + \tau - \gamma_1 \varepsilon_e - \varepsilon_u(1 + \gamma_1) \\ oe_a = \gamma_0 + \gamma_1 p + \varphi, \end{gathered} \tag{A5}$$

with the composite error term $\varphi = \tau - \gamma_1 \varepsilon_e - \varepsilon_u(1 + \gamma_1)$.

It is straightforward to see that OLS leads to biased estimates of $\gamma_1$ because the known and unknown luck components are part of performance and the composite error term. The simple OLS bias is equal to $\frac{Cov(p,\varphi)}{Var(p)}$. To obtain a clearer picture of its direction, we decompose and rearrange the bias term as follows:

$$\begin{gathered} \frac{Cov((s^* + \varepsilon_k + \varepsilon_u), (\tau - \gamma_1 \varepsilon_k - \varepsilon_u(1 + \gamma_1)))}{Var(p)} \\ -\gamma_1 \frac{Var(\varepsilon_k)}{Var(p)} - (1 + \gamma_1) \frac{Var(\varepsilon_u)}{Var(p)}. \end{gathered} \tag{A6}$$

The two terms of Equation (A6) allow us to understand the direction of the bias in eight scenarios, which depend on the role of known and unknown measurement error. In particular, we consider the cases with 1) no measurement error, 2) only known measurement error, 3) only unknown measurement error, and 4) known and unknown measurement error.

For each of these cases, we consider two cases of $\gamma_1$: first, the case where -1< $\gamma_1 < 0$. This case reflects a situation in which the relationship between skill and overconfidence is negative, as suggested by the Dunning–Kruger effect, but larger than -1, because $\gamma_1 \leq$ -1 would mean that self-assessed skill stays constant or even declines with actual skill (see Section 3.3). The second case is $\gamma_1$= 0. This case reflects a situation in which there is no relationship between skill and overconfidence and the empirical negative relationship between performance and overestimation is a statistical artifact driven by measurement error.

Table A1 shows the directions of the bias in the resulting eight scenarios. If performance measures skill perfectly ($Var(\varepsilon_k) = 0$ and ($Var(\varepsilon_u) = 0$), both terms in Equation (A6) are equal to zero and OLS estimates are unbiased. If performance measures skill with some error but this error is perfectly known ($Var(\varepsilon_k) > 0$ and $Var(\varepsilon_u) = 0$), the second bias term is equal to zero. This means that OLS is unbiased if there is no relationship between skill and overconfidence ($\gamma_1$= 0) and underestimates the Dunning–Kruger effect when the relationship is negative ($\gamma_1 < 0$). If performance measures skill with some error but this error is completely unknown ($Var(\varepsilon_k) = 0$ and $Var(\varepsilon_u) > 0$), the first bias term is zero and OLS leads to a negative bias if the Dunning–Kruger effect exists (-1 $\leq \gamma_1 < 0$) and if it does not ($\gamma_1$= 0). Finally, the most realistic scenario is that performance measures skill with some error and this error is known only to some extent after the test ($Var(\varepsilon_k) > 0$ and $Var(\varepsilon_u) > 0$). In this case, the direction of the bias again depends on the true relationship between skill and overconfidence. If the Dunning–Kruger effect is correct, the bias may be positive or negative, depending on whether the positive bias of the first term is larger than the negative bias of the second term. If the Dunning–Kruger effect is incorrect and there is no relationship between skill and overconfidence ($\gamma_1$= 0), OLS leads to an overestimation of the Dunning–Kruger effect.

**Table A1: OLS Bias when Eliciting Performance Expectations after the Test**

| | (1) | (2) |
|---|---|---|
| Measurement error | $-1 < \gamma_1 < 0$ | $\gamma_1 = 0$ |
| $Var(\varepsilon_k) = 0$ & $Var(\varepsilon_u) = 0$ | unbiased | unbiased |
| $Var(\varepsilon_k) > 0$ & $Var(\varepsilon_u) = 0$ | positive bias (underestimation of DKE) | unbiased |
| $Var(\varepsilon_k) = 0$ & $Var(\varepsilon_u) > 0$ | negative bias (overestimation of DKE) | negative bias (overestimation of DKE) |
| $Var(\varepsilon_k) > 0$ & $Var(\varepsilon_u) > 0$ | ? | negative bias (overestimation of DKE) |

**Note:** This table shows the directions of the bias for the eight scenarios that depend on $\gamma_1$, $Var(\varepsilon_k) = 0$, and $Var(\varepsilon_u)$. The Dunning–Kruger effect (DKE) states that $\gamma_1$is negative, so a positive bias is an underestimation of the DKE and a negative bias is an overestimation of the DKE.

How does the magnitude of this bias compare to the bias when performance is elicited before the test? For comparison, recall that when performance is elicited before the test, the bias is equal to $-(1 + \beta_1)\frac{Var(\varepsilon)}{Var(p)}$. This term is analogous to the second term of Equation (A6). In the most realistic scenario, where luck is known to some extent after the test, the OLS bias is likely to be less negative (or more positive) for two reasons. First, in the case that the Dunning–Kruger effect does not hold, the bias is smaller, because the unknown part of the measurement error is only a subset of the overall measurement error and $Var(\varepsilon_u) < Var(\varepsilon)$. Second, if the Dunning–Kruger effect holds, the first term of Equation (A6) leads to a positive bias, counterbalancing the negative bias of the second term.

We conclude that, because test takers are unlikely to perfectly know their luck after the test, we cannot rule out that the observed negative relationship between performance and overestimation elicited after the test is a statistical artifact. The magnitude of the overall bias is, however, likely smaller.

How does eliciting performance expectations after the test affect IV estimates? Analogous to Equation (10), the following equation shows the decomposed IV estimator of $\gamma_1$:

$$\frac{Cov(oe_a, p_2)}{Cov(p, p_2)} \qquad \text{(A7)}$$

$$\frac{Cov(oc_a, s^*) + Cov(oc_a, \varepsilon_2) + Cov(\varepsilon_u, s^*) + Cov(\varepsilon_u, \varepsilon_2)}{Var(s^*) + Cov(s^*, \varepsilon_2) + Cov(\varepsilon_k, s^*) + Cov(\varepsilon_k, \varepsilon_2) + Cov(\varepsilon_u, s^*) + Cov(\varepsilon_u, \varepsilon_2)}.$$

Recall that the original performance $p$ measures skill with a classical measurement error, which implies that skill is unrelated to the known and unknown luck of this performance ($Cov(\varepsilon_u, s^*) = Cov(k, s^*) = 0$). If this is the case, we can see from Equation (A7) that IV corrects for bias caused by measurement error if the luck portion of the second skill measure is uncorrelated with skill ($Cov(s^*, \varepsilon_2) = 0$), overconfidence ($Cov(oc_a, \varepsilon_2) = 0$), and the known and unknown luck parts of the original performance ($Cov(\varepsilon_k, \varepsilon_2) = Cov(\varepsilon_u, \varepsilon_2) = 0$). In principle, these cases should be equally plausible, whether performance expectations are elicited before or after the test.

**A2: Questionnaire**

Page 1 of the questionnaire starts below:

---

Dear student,

I am [anonymized]. My research concerns the relation between grade expectations and realised grades.
I would like to ask you for **your expectations of your grade in the [course name] exam and your participation grade**. Please give your best estimates. You can enter three lotteries if your estimates are close to your actual results. In each lottery you can win one of three VVV vouchers worth €20. In total, you can win VVV vouchers of €60.
At the end of the survey, you will be asked to enter your student ID. The ID is required to compare your estimates with your actual results. If you win one of the lotteries, the ID will be used to look up your email so that I can inform you about your win.
**I will treat this information confidentially and ensure your anonymity.** No individual information will be passed on to anybody (not even your tutor or course coordinator). I will also not report any information which can be used to identify you.
If you have any questions, please feel free to contact me via: [anonymized]
Thank you for your cooperation!
[anonymized]

This is how the lotteries are going to work:

Lottery 1: If your exam grade (in your first attempt) is within 0.25 points of your expected grade, you enter a lottery in which two winners are randomly drawn. If you do not attend the first sit, your second sit grade is considered for the lottery. Each winner will receive a VVV voucher worth €20.

Lottery 2: I calculate the actual percentile of your exam grade compared to the exam grades of the first attempts of all students in this course. If your final exam grade is in your expected percentile range, you enter a lottery in which two winners are randomly drawn. Each winner will receive a VVV voucher worth €20.

Lottery 3: If your actual participation grade is within 0.25 points of your expected participation grade, you enter a lottery in which we randomly draw two winners, who will receive a VVV voucher worth €20.]

**Questionnaire Grade Expectations - Course [course name]**
**********

1. **Which grade do you expect to get in the exam of the course [course name]?**

*If you do NOT intend to attend the first sit, please state your expectations for the second sit (resit).*

- I expect to get a __ __.__ __ in the exam. [0.00-10.00]

**2. Please indicate in which <u>percentile range</u> you expect your exam grade to be in?**

*The percentile shows the percentage of students in this course which have a lower exam grade (in their first attempt) than you. High values mean high exam grades compared to the exam grades of the other students in this course.*

***Please mark your expected percentile range with an X.***

| | 1-10% | 11%-20% | 21%-30% | 31%-40% | 41%-50% | 51%-60% | 61%-70% | 71%-80% | 81%-90% | 91%-100% |
|---|---|---|---|---|---|---|---|---|---|---|
| Your percentile: | | | | | | | | | | |
| | *Worst 10%* | | | | | | | | | *Best 10%* |

**3. Which <u>participation grade</u> do you expect to get in this course?**

*Please state your guess rounded to the next quarter point so that it ends with .00, .25, .50 or .75.*

- I expect to get a__ __ .__ __as participation grade. [0.00-10.00]

**4. Do you consider <u>failing on purpose in the first sit of the exam</u> in this course – either by not attending or by handing in an incomplete exam – in order to get a higher grade in the second sit?**

☐ Yes ☐ No

**5. What is your gender?**

☐ Male ☐ Female

**6. What is your student ID?**

- **ID_____**

---

Please fold this page in half after filling it out.